DOCUMENT RESUME

ED 248 262 TM 840 565

AUTHOR Bangert-Drowns, Robert L.
TITLE Developments in Meta-Analysis: A Review of Five Methods.
INSTITUTION Michigan Univ., Ann Arbor. Center for Research on Learning and Teaching.
REPORT NO CRLT-36
PUB DATE Jun 84
NOTE 69p.
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Estimation (Mathematics); Literature Reviews; *Meta Analysis; Probability; *Research Methodology; Research Utilization; Scientific Research; Statistical Analysis

ABSTRACT

It is easy to observe that meta-analysis is quickly establishing itself as useful tool of the social sciences. Perusal of representative journals confirms that meta-analysis has been applied in various ways to diverse literatures. It is imperative, therefore, that reviewers, publishers, consumers, and critics of these reviews be best informed about the method. It is especially important to clarify exactly what the term "meta-analysis" refers to. This article proposes a clarification in two ways. First, meta-analysis is compared to and distinguished from other methods of research integration that preceded it. Second, five different types of meta-analytic method are distinguished by their purposes (for scientific criticism or for literature review) or by their methods (using combined probability, using tests of homogeneity, or using estimates of population variation). (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED248262

Developments in Meta-analysis:

A Review of Five Methods

Robert L. Bangert-Drowns

Center for Research on Learning and Teaching

University of Michigan

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Bangert-Drowns

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CRLT Report Number 36

June 1984

Center for Research on Learning and Teaching

109 East Madison

Ann Arbor, Michigan 48109

TM 840 505

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Abstract

It is easy to observe that meta-analysis is quickly establishing itself as a useful tool of the social sciences. Perusal of representative journals confirms that meta-analysis has been applied in various ways to diverse literatures. It is imperative, therefore, that reviewers, publishers, consumers, and critics of these reviews be best informed about the method. It is especially important to clarify exactly what the term "meta-analysis" refers to. This article proposes a clarification in two ways. First, meta-analysis is compared to and distinguished from other methods of research integration that preceded it. Second, five different types of meta-analytic method are distinguished.

Table of Contents

Abstract i

Introduction 1

Forms of Research Integration 2

Precedents for Meta-analysis 11

The Consensus Regarding Meta-analytic Method 18

Divergence in Meta-analytic Method 22

For Scientific Criticism 23

For Literature Review 27

Using Combined Probability 31

Using Tests of Homogeneity 37

Using Estimates of Population Variation 44

Discussion 48

Conclusion 57

Bibliography 58

An Exposition of Developments in Meta-analytic Method

When Glass (1976) first coined the term *meta-analysis*, he meant the term to refer to no new or specific technique. Instead, Glass advocated the application of familiar experimental methodologies to summarize research findings in the social sciences. As he put it later, "The approach to research integration referred to as 'meta-analysis' is nothing more than the attitude of data analysis applied to quantitative summaries of individual experiments...it is not a technique; rather it is a perspective that uses many techniques of measurement and statistical analysis" (Glass, McGaw, & Smith, 1981, p. 21).

Since then, there has been some disagreement over what exactly constitutes meta-analysis. For example, Hunter, Schmidt, and Jackson (1983) have referred to their meta-analytic work as "state-of-the-art meta-analysis...the most complete meta-analysis procedure now known" (p. 140). Hedges and Olkin (1980) regard Glass's original formulation as outdated and present what they believe to be a more technically adequate form. Glass has recently restated his confidence in his original formulation of meta-analysis and called for greater specificity in the use of the term (Glass & Kleigl, 1983).

For the reader who knows little about meta-analysis, these disagreements can only further obscure what was ironically intended to be a tool of clarification. Readers, editors, and research reviewers need to be better informed if they are to be intelligent consumers and critics of meta-analytic reviews. Meta-analysis is becoming an increasingly commonplace tool of social science. Kulik (1984) estimates

that about 320 meta-analyses have already been conducted. The need for more information about this method is, therefore, also growing. This article will provide some of the needed information. First, meta-analysis will be compared with other forms of research integration. Second, five types of meta-analytic method will be distinguished.

Forms of Research Integration

The research of the social scientist is rarely as elegant as the critical experiments of physics and chemistry. Especially in areas of applied research, studies investigating identical problems may differ slightly in the population sampled, the treatment or dependent measure used, the methodological design, the duration of the evaluation, the experimental setting, and the researcher's inadvertent influences. Such studies cannot be strictly considered replications. Though their outcomes may be highly correlated, they cannot be expected to duplicate each other (Kraemer & Andrews, 1982). As a result, the numbers of related, yet nonreplicative studies are so staggering that unaided interpretation of their findings is impossible.

Economic pressures make interpretation of this confusion more urgent. Funding sources generally expect that the research they sponsor will return some tangible product, either in the form of some generally applicable tool or some specific measure that can guide future policy determinations. If researchers in a given discipline can only deliver contradictions or findings relevant to only a small sample, funders will become more skeptical of the ways their money is being used.

Research integration is one response to these pressing concerns. The most familiar form of research integration is the traditional

narrative review. The narrative reviewer collects some samples of studies and uses his or her expertise to offer speculative explanations for the discrepancy in study outcomes.

The meta-analytic solution is an alternative to this narrative review. Just as every study describes regular patterns discoverable among complex data, so each study or study outcome is one of a great new collection of data points. This collection of study outcomes can, in its turn, be statistically described.

Many observers suggest that research integration could be improved if it were more quantitative, more statistically sophisticated, and as methodologically rigorous as other scientific study. While primary research is subject to critical evaluation by all kinds of consensual standards for theoretical, methodological, statistical, and interpretative adequacy, "there are no consensually accepted rules for the scholar seeking to determine the convergences and disagreements within the literature" (Fiske, 1983, p. 65). It is natural that research review would be expected to operate by the same demanding standards as the research it reviews (Glass, McGaw, & Smith, 1981). Jackson (1980) and Cooper (1980) agree that the process of research review should duplicate the five steps used generally in scientific research: problem formulation, data collection, evaluation of data points, analysis and interpretation, and presentation of results.

Practically speaking, the greatest difficulty faced by the traditional reviewer is finding a satisfactory method of managing the large number of studies in an area. As Hunter, Schmidt, and Jackson (1983) put it, the reviewer using an informal narrative review is left with three alternatives. The first is a pedestrian listing of all the

studies amassed, highlighting special independent variables and outcomes for each. This approach produces little more than an annotated bibliography and leaves the responsibility of integration and interpretation to the reader. The brief descriptions of each study may be so sketchy, however, that significant relationships between study features and outcomes are masked. Secondly, the reviewer may restrict the studies being considered to a more manageable number. While integration and interpretation are easier, such restrictions are generally unsystematic. Such a restricted collection may indicate likely outcomes of other studies, but it may also reflect the reviewer's selection biases.

The third alternative for the narrative reviewer is to include all the relevant studies, integrate them, and draw conclusions. Hunter, Schmidt, and Jackson (1983) argue that, without statistical aid, some quantitative information that could have been preserved will be necessarily dropped. Glass, McGaw, and Smith (1981) liken this third approach to a researcher attempting to summarize hundreds of test scores without statistical analysis.

Jackson (1980) has gathered empirical evidence to demonstrate the need for a "meta-analytic attitude" in research integration. He randomly collected 36 narrative review articles from several respected journals. His observations of these articles provide an important criticism of the narrative style of review as it is actually practiced. Only two of the thirty-six reviews made critical usage of previously published reviews. Reviewers almost never reported the search procedures for the studies they examined nor did they explain the criteria by which they selected the final group of studies. Only one of

the thirty-six reviews discussed the full set of studies which the author had located.

In regard to representing the outcomes of the studies, Jackson found that only four of the 36 randomly selected reviews attempted to analyze their findings in terms of significance and direction. Since so few narrative reviews used a statistical examination of the research outcomes, reviewers could not accurately assess the probability that a particular configuration of primary findings could occur by chance. Obviously, the degree to which a particular configuration of research findings is likely to randomly occur should bear directly on the confidence that a reviewer can assign to his or her conclusions. Jackson also found that only three reviews looked for relationships between study features and study outcomes. In attempting to draw broad generalizations about specific independent variables, reviewers typically ignored the possibility that additional situational factors were moderating the outcomes of experimental applications.

The narrative review is typically faulted for unnecessary loss of information, especially quantitative information, in integrating studies and for bias in sampling the available literature. Some observers, however, are quick to point out the natural similarities between narrative review and more quantitative meta-analytic review. Cook and Leviton (1980) argue that the flaws observed by Jackson are not inherent in traditional reviews but indicate only that narrative review is being poorly done. Meta-analysts, they continue, must be as concerned as narrative reviewers about bias in sampling, methods of representing study outcomes, and interpretation. Not surprisingly, meta-analyses can often suffer the same problems as narrative reviews.

The methodology of research integration, therefore, is best seen not so much as dichotomous, that is, not as either narrative or meta-analytic review, but as a continuum whose poles are roughly represented by the two kinds of review. On the one hand, there is the informal collection of studies whose outcomes are merely listed or loosely gathered in impressionistic observations. Moving toward the other pole, review procedures are more explicitly defined. Sampling procedures are specified and methods for representing individual studies in the review are increasingly enhanced to avoid bias and the loss of quantitative information. Finally, and most importantly, conclusions are drawn from the collection of studies only after an analysis of the study outcomes in light of sampling theory. It is this tendency toward greater explicitness, specificity, replicability, quantification, and statistical analysis that characterizes the meta-analytic approach to research integration. Meta-analysis can improve research review even though it may share some of the limitations of narrative forms.

Mintz (1983) has described this continuum of research integration most recently. "As the review process progresses from the descriptive narrative summary to the abstract heights of the multiple regression analysis, a series of steps is taken by the reviewer. Each step involves increased quantification and abstraction" (p. 71). A similar observation was made at least as early as 1971 by Light and Smith.

In their discussion of synthesizing techniques that are pre-meta-analytic, Light and Smith described four methods of increasing sophistication. In the simplest synthesis, the reviewer simply lists all the variables found to effect a dependent variable. The frequency, statistical significance, or degree of effect is not accounted for here.

In the second form, the reviewer chooses two or three studies determined to be the "best" in some area. The determination of "best" is typically made on methodological grounds. By the third method, collected studies are ordered according to their outcomes. The study producing the median value is assumed to represent the set. The fourth approach is called *vote-counting*, determining the frequency of studies presenting outcomes of differing direction and significance. These methods represent a continuum increasing in explication, quantification, and statistical sophistication.

Since the vote-counting method of research review is generally considered the most sophisticated form of narrative research integration, it is worth examining more closely. This method is also known as the "box score review" or modal categorization.

A reviewer using vote-counting will divide a collection of studies into three smaller sets: those significantly favoring the control group, those significantly favoring the experimental group, and those studies with nonsignificant outcomes. (Some reviewers will divide the nonsignificant findings into two smaller groups, either favoring the control or the experimental treatment.) The modal categorization is taken as the best representation of the outcomes of this research area. Outcomes of lesser frequency are attributed to chance or undetected experimental error.

The vote-counting method is a more sophisticated form of the narrative review because it shows a primitive sensitivity to the quantification of outcomes and to the operation of chance in these outcomes. By counting statistically significant findings, this method emphasizes those studies whose outcomes are most reliable and whose

treatment effects are largest, most positive or negative. By counting the frequency of such outcomes, the reviewer has a crude idea of how reliable a particular finding is throughout a body of research.

It is not hard to imagine, however, collections of studies which would be misrepresented by this method. For example, a researcher might gather 200 studies of the effectiveness of psychotherapy. Of these studies, 95 significantly favor the control group, 70 significantly favor the use of psychotherapy, and 35 report nonsignificant findings. The most reliable conclusion that the vote-counting reviewer can draw is that psychotherapy is ineffective since more statistically significant findings favor the control condition.

There are several hazards in this interpretation, however. First, vote-counting is typically uni-variable, counting significant differences for or against one specific treatment. Vote-counting can only examine the independent variable for which significance levels are computed and therefore "almost invariably focuses upon the main effect of each variable in isolation from any other potentially interactive factors" (Light & Smith, 1971, p. 437). In our hypothetical case, the distribution of the findings is bimodal. This suggests the influence of an additional intervening variable which cannot be examined in the framework of the univariable vote-count.

Secondly, the vote-counting method is systematically biased to reflect the outcomes of studies with larger sample. This is because statistical significance does not only reflect the size of a treatment effect, but also the size of a study's sample. For example, the 95 psychotherapy studies favoring the control group may actually show very small differences between experimental and control treatments. However,

even these small treatment effects could be statistically significant if their samples are large. The remaining 105 studies may all reflect very large positive psychotherapy effects, only 35 of these outcomes are nonsignificant because the sample sizes of these studies are small. By just counting statistical significance, modal categorization did not indicate a consistently large treatment effect.

Vote-counting will only reliably represent the size of a treatment's effect if the collected studies have the same sample sizes and a unimodal distribution reflecting one population. Even under these circumstances, a modal category still may not accurately represent a population treatment effect. The distribution of study outcomes must still be considered in the light of sampling error.

The benefits of more statistically rigorous integrations of study outcomes seems intuitively obvious. Cooper and Rosenthal (1980) offer empirical evidence to support this intuition. These researchers created two groups, each of graduate students and faculty volunteers. All subjects were given seven studies on sex differences in persistence. In one group subjects were asked to summarize the studies in any way they believed appropriate. The second group was given a specific procedure to estimate the probability of the combined outcome. Sixteen of the twenty subjects using informal review concluded that there was no relationship between sex differences and persistence. Conversely, only six of twenty subjects using the statistical review drew the same conclusion. The latter group also reported a greater sense of certainty regarding their conclusions than did the former group. The greater confidence, accuracy, and replicability of the statistical review

suggests the advantages of a verifiable, quantitative, integrative technique.

Methods of research integration can be roughly categorized as meta-analytic or narrative, depending on the degree to which they use explicitly stated and unbiased sampling techniques, quantitative or quasi-quantitative representation for study outcomes and study features, and statistically sophisticated methods for integration and analysis. The differences between these two categories may be more continuous than dichotomous. In addition, the meta-analytic view is not intended as a replacement for traditional review, but as a complementary integrative technique. Cook and Leviton (1980) warn that meta-analysis is inappropriate with a small number of heterogeneous studies. "When studies are heterogeneous, we must have larger samples to be confident that the studies are representative of an underlying distribution" (p. 461). Glass and Kliegl (1983) argue that the personal assessments of a seasoned expert are still to be valued and that reviews investigating a few studies in great detail can still provide useful information for policy theory and research. The meta-analytic attitude is thus limited in appropriateness.

Four purposes for research review can be distinguished (Jackson, 1980; Cooper, 1980). One goal of review is to identify and assess new developments in a field. A second aim is to use empirical evidence to highlight, illustrate, or assess a particular theory or to tentatively propose new theoretical frameworks. Third, a reviewer can organize knowledge from divergent lines of research. All three of these purposes could best be served by more discursive reviews. But a researcher may

also want to integrate knowledge from convergent research efforts. It is to this fourth purpose that meta-analysis is best applied.

Precedents for Meta-analysis

Glass has insisted that his formulation of meta-analysis was not a major break from past review efforts (Glass, McGaw, & Smith, 1981, p. 24). In fact, he cites two reviews that are exemplary of earlier applications of meta-analytic ingenuity. Underwood (1957) located 16 experiments measuring the effects of interference on recall. Graphing the apparently divergent outcomes of the 16 studies, Underwood discovered a significant relationship between prior lists learned and recall ability. This relationship was not detected in any individual study. In the second case, Sudman and Bradburn (1974) gathered hundreds of studies investigating response effects in survey research. These reviewers applied statistical methods to assess the influence of survey design and use on survey results.

These two studies indicate that the meta-analytic attitude was already, before Glass, a natural development of the desire for greater quantitative and statistical rigor in organizing diverse findings in social science research. Other well-known examples are not hard to find. Bloom (1964) aggregated correlation coefficients to summarize evidence of stability and change in behavior. Cartwright (1971) quantified and cumulated study outcomes in his review to show that group influence is trivial in risky shift research. Erlenmeyer-Kimling and Jarvik (1963) aggregated more than 30,000 correlation pairings from 52 studies to discover orderly relationships between intelligence and genetic endowment.

One can trace meta-analysis to earlier roots (Hedges & Olkin, 1982). In the first third of this century, agricultural experimenters assembled a large body of studies bearing on identical questions. In analogous ways, interpretation of these study findings was hampered as interpretation in the social sciences is today.

> First, the effect of most factors (fertilizers, varieties, etc.) varies considerably from place to place and from year to year, owing to differences in soil, agronomic practices, climatic conditions and other variations in environment. Consequently the results obtained at a single place and in a single year, however accurate in themselves, are of limited utility...Secondly, the execution of any large-scale agricultural research demands an area of land for experiment which is not usually available at a single experimental station...The agricultural experimenter is thus frequently confronted with the results of a set of experiments on the same problem. (Yates & Cochran, 1938, p. 556)

Tippitt (1931) answered the difficulty by supplying methods to test the statistical significance of results combined from separate experiments. Other researchers, such as Fisher (1932) and Pearson (1933, 1938), followed in this tradition. They sought ways to combine probability values from tests of significance.

Cochran (1937) and Yates and Cochran (1938) set out to answer the difficulty in a different way. They developed methods to estimate the mean effect and variability of a particular treatment examined at different agricultural centers. Furthermore, Cochran (1943) discussed the importance of comparing different dependent measure criteria for compatibility and relative accuracy before combining study findings.

Wilkinson (1951) made an early proposal to aid the interpretation of multiple outcomes from single studies in the social sciences. He observed that, since study results are either significant or nonsignificant, the binomial distribution could be used to determine the expected number of significant results given a true null hypothesis. This estimate of expected studies could be tested against the actual number of significant findings as a check on the assumption about the null hypothesis. Jones and Fiske (1953) went further to import the meta-analytic attitude of agricultural research to the social sciences. They extended Wilkinson's proposition and argued that Fisher's and Pearson's method of combining probabilities could be used to combine findings in education and psychology. Mosteller and Bush (1954) also established combined probabilities as a useful tool in social science integration.

As early as the 1950s, therefore, combining probabilities was advocated as a reasonable way of statistically integrating social science research. However, no integrative technique similar to Cochran's was formally developed for the social sciences. That is, no one advocated a coherent method to estimate the mean and variation of a social or psychological treatment. Two remaining methodological problems blocked this development.

Dissimilar dependent measures and variations in study features presented difficult obstacles to the translation of the meta-analytic attitude into a coherent meta-analytic method. Though a group of studies might investigate the same question with similar methods, the reviewer could not be sure to what degree the combined results were distorted due to differences in the criteria scale used and differences

in study features. Two influential articles offered solutions to these problems and in many ways anticipated what would become modern meta-analytic method.

In his well-known article "The statistical power of abnormal-social psychological research," Cohen (1962) indirectly, but nonetheless potently, spoke to some of the gaps in meta-analytic method. His concern was the absence of power analysis in the social sciences. As a study is more capable of detecting a false null hypothesis, it is deemed more statistically powerful. Social scientific convention, said Cohen, seems more concerned with the rejection of a true null hypothesis (i.e., statistical significance) than with accepting a false one. He called for a greater balance. In developing this argument, Cohen not only clarified methods for the calculation of effect size, but his empirical investigation formally resembled what Glass would later call meta-analysis.

To demonstrate the inadequate attention given to power analysis, Cohen set out to review the power of the 78 articles appearing in the Journal of Abnormal and Social Psychology, 1960, 61. Cohen decided to make three calculations of statistical power for each article, each calculation corresponding to hypothetical population effects of different degrees: low, medium, or large. Given such calculations, Cohen showed that the 78 articles were poorly designed to detect false null hypotheses when effect sizes were medium or low.

To the historian of research integration, Cohen's review made three important contributions to the development of meta-analytic method. First, Cohen showed a way of statistically summarizing a diverse literature without taking recourse to counting or combining significance

levels. The articles reviewed were methodologically varied, sharing only two features: they were published in a particular volume of a particular journal and they tested the probability of null hypotheses. Each study was represented in the review by three power estimates. These estimates were summarized in terms of central tendency (mean and median) and variation.

Secondly, since effect size is a factor in determining a study's power, any call to power analysis must also be an encouragement for an analysis of effect size. Cohen did not calculate the effect size for each study, a procedure that is common in modern meta-analysis. He did, however, underscore the convenience of the use of effect size in statistical research integration. "In the present study, the problem [of effect size estimation] was further complicated by the need to answer it for diverse content areas, utilizing a large variety of dependent variables and many different types of statistical tests." His solution came easily. "For each type of statistical test where it was necessary, size of effect was expressed quantitatively in terms not dependent on the specific metric of the variable(s) involved" (p. 446). Cohen offered a variety of formulas for effect size calculations and estimated power for each study by simple algebraic manipulations of hypothetical effects.

Third, Cohen suggested a rudimentary form by which the relations between study features and study findings could be examined. Cohen realized that the power of major statistical tests, tests on the central hypotheses of the article, may be greater than the power of more peripheral statistical tests. He classified these separately, performing his analysis on the major tests, then analyzing the power of

the peripheral tests to see if there were differences between the classes. In this way, the influence of a study feature on findings was empirically assessed.

A second influential article speaks more directly to the development of meta-analytic method. In 1971, Light and Smith published their article "Accumulating evidence: Procedures for resolving contradictions among research studies." Like most researchers, these authors were concerned with how to translate the contradictions of experimental findings into policy recommendation and reliable scientific knowledge. By their standards, the pre-meta-analytic or narrative methods of research integration were inadequate. Their most sophisticated form, the vote-counting method, implicitly assumes that each study is a random sample taken from a common population. Attempting to synthesize research findings under such an assumption may only mask real and important differences among clusters or naturally aggregated subpopulations of the overall population.

Light and Smith called their alternative *cluster analysis*. Cluster analysis involves five steps. First, studies are accepted to cluster analysis only if they meet certain criteria. The subjects in each study must come from an identified population dependent, independent variables must be measured in the same way, and the instrumentation and quality of the studies must be roughly equivalent. In a word, studies used in cluster analysis must be methodological replications, while variation among subjects remains.

Second, the reviewer must gather the original data reflected in the studies' findings. Third, the original data are regrouped into natural aggregations. The development of aggregation is the judgement of the

reviewer determined by balancing his or her purposes with the intentions of the primary researchers. Clusters may consist of schools, classes, neighborhoods, or some other natural aggregation.

Once the original data are regrouped into clusters, differences among clusters are investigated to see if they adequately explain apparent differences in the original studies. Clusters may differ in their means, in their variances, in the relations between independent and dependent variables, in subject/treatment interactions, and in cluster effects (i.e., subject/cluster interactions). All five of these potential differences must be considered.

Finally, in those cases where no differences are found, clusters can be combined to increase sample sizes. At the end of a cluster analysis, a reviewer should be in a position to draw conclusions from larger samples or to attribute differences in study findings to identifiable differences among the clusters. Of course, such an analysis may also fail and the variation in the original studies be found to be unexplainable.

In many ways, cluster analysis is distinguishable from the meta-analytic tradition. Cohen (1962) was willing to quantitatively represent each study in his review of power analysis and perform analyses on these statistical representations. This resembles modern meta-analytic procedure. Light and Smith, on the other hand, relied on original data. In this way, cluster analysis is more a form of secondary analysis, the reanalysis of original data, than a form of meta-analysis and is impossible to do in the numerous cases where original data is unavailable.

Yet Light and Smith (1971) have made an important contribution to the maturation of meta-analysis. Their interest was, after all, in the synthesis of diverse study outcomes. They reaffirmed the importance of finding a statistically sensible way to resolve apparent contradictions in research findings. More importantly, their work underscores the necessity of the search for subject/setting interactions as an essential focus of any research integration and an essential means for explaining variation in study findings.

The Consensus Regarding Meta-analytic Method

It is important to distinguish the meta-analytic perspective or meta-analytic attitude from meta-analytic method. The meta-analytic perspective is the pervasive and growing *weltgeist* in the social sciences which seeks to bring statistical sophistication to the process of research integration. This movement is borne of economic pressures and a respect for scientific rigor.

Meta-analytic method is another matter. This is the specific technique the reviewer uses to produce a quantitative research integration. The meta-analytic attitude has existed in the social sciences for some time and, in itself, is generally regarded favorably. Meta-analytic method has only recently been discussed with much seriousness and it has attracted a fair number of devotees as well as ardent critics.

Glass originally portrayed meta-analysis as method-free; that is, the meta-analyst is free to use any statistical method at hand so long as the application is reasonable to the purposes of the review and does not misrepresent the outcomes of the studies being considered. Yet any

superficial perusal of a group of meta-analyses would reveal a fair degree of similarity in methodology. Even Glass has more recently called for a more methodologically specific use of the term meta-analysis (Glass & Kleigl, 1983). There is sufficient consensus to begin a tentative definition of meta-analysis as a method.

Meta-analysis is the quantitative review of an explicitly defined body of research. Its general purpose is to describe and explain variability in the outcomes of that body of research. No particular meta-analysis should be considered the last word in any research field. Meta-analytic findings, like all findings in social science research, are interpretations of data seen from a particular researcher's perspective (Glass & Kliegl, 1983; Wilson & Rochman, 1983). There are numerous judgement points in meta-analysis where the reviewer's biases can influence his or her conclusions and meta-analysis does not remove all subjectivity from research integration. Rather, it permits greater precision in measuring research outcomes. Secondly, meta-analysis is sufficiently explicit to permit easy identification of the review's limitations by a reader and replication by other reviewers. Finally, good meta-analysis will indicate areas where continuing research would be most fruitful (Fiske, 1983).

As a quantitative review, meta-analysis is distinguishable from primary research, secondary research (or secondary analysis), and narrative review (Glass, 1976). Primary research is the original gathering and analysis of information about treatment effects on a sample. Secondary research or secondary analysis is the re-analysis of original data in the hope of correcting some suspected flaw or inadequacy in a primary analysis. Narrative review is the informal

critical evaluation of a group of studies. Meta-analysis can be distinguished from these at the levels of sample selection, data collection, and analysis.

The sample chosen for meta-analytic review should roughly represent all the relevant studies bearing on a specifically defined question. Thus, while some studies may escape inclusion because of irretrievability or a reviewer's oversight, the meta-analytic sample is not representative or random but a close approximation to the existing population. There is some debate about whether the meta-analyst should apply selection criteria to exclude studies considered methodologically inadequate. Glass, McGaw, and Smith (1981) have argued that methodologically inferior studies be included in meta-analysis so that the effects of various flaws can be empirically determined. Others are more concerned about how the inclusion of inferior research may effect meta-analytic findings and recommend statistical adjustments or exclusion (e.g., Hunter, Schmidt, & Jackson, 1980; Bangert, Kulik, & Kulik, 1983). In either case, the meta-analyst must specifically describe the search procedures (what databases were searched, how were they searched) and any selection criteria that were used.

Meta-analyses examine large numbers of studies. Statistical analysis of small groups of findings will produce unreliable results. Even if a large number of outcomes are taken from a small sample of studies, the reviewer has the interpretative problem of contending with the lack of independence among outcomes. A narrative review might be better suited to groups of studies numbering less than ten to fifteen (Cook & Leviton, 1980).

Every selected study is represented in the meta-analysis, though studies may be given differential weight depending on their methodological quality or relevance to the question under review. Each study is represented in two ways. Study outcomes are given at least one quantitative representation. Typically study outcomes are translated into Cohen's *d* or Glass's effect size, the difference between experimental and control group means in units of standard deviation. Recently, other representations for outcomes have been recommended (Hedges, 1980; Rosenthal & Rubin, 1982a; Kraemer & Andrews, 1982), but these have not gained general acceptance.

In addition to their outcomes, studies are represented according to their features. The reviewer defines a set of study characteristics which possibly bear relationship to study outcome, for example, random vs. nonrandom subject assignment, duration of treatment, and variations in treatment application. For each characteristic, the study is given a quantitative or quasi-quantitative categorical coding.

When coding is finished and the data collected, analysis can begin. The central tendency and variability of the study are measured. The essential work of meta-analysis, attempting to explain variation in study outcomes, is accomplished by testing the effects of study features on study outcomes. Typically parametric tests (especially t-tests, ANOVA, and multiple regression) are used to examine these relationships. Hedges (1982a, 1982b, 1983), Hunter, Schmidt, and Jackson (1983), and Rosenthal and Rubin (1982b) have recommended statistical tests which they believe are more appropriate to meta-analytic purposes, but these are not in general use.

Divergence in Meta-analytic Method

After recognizing the present degree of consensus in the evolution of meta-analytic method, the reader should be immediately alerted to the degree of divergence. Glass and Kleigl (1983) called for a more specific use of the term *meta-analysis* precisely because there is confusion over what the term labels. The term has been applied to Light and Smith's cluster analysis (Cook & Leviten, 1980), to a review that coded study features but did not statistically analyze outcomes (Dickson, in press), to vote-counting (Strube & Hartman, 1983), and to a review of only 12 studies (Ide, Parkersen, Haertel, & Walberg, 1980). These reviews resemble meta-analysis, but by stricter definition fall outside the consensual definition.

Yet there are differences even among methods generally agreed to be meta-analytic. In 1976, the same year that Glass first coined the term meta-analysis, Rosenthal published his book Experimenter Effects in Behavioral Research and Schmidt and Hunter were working on a validity generalization technique. These three concurrent efforts have resulted in three distinguishable meta-analytic methods. Since the late 1970s, two additional coherent methods have been formulated as elaborations of Glass's approach, one by Kulik and the other by Hedges and Rosenthal and Rubin. No understanding of the meta-analytic method would be complete without reviewing the differences among these five approaches. They are distinguishable on four factors: purpose, consideration of study features, representation of treatment effect, and unit of analysis.

Meta-analysis as a Tool of Scientific Criticism

Scientific literature does not just record the measured effects of controlled treatments. It is also an enduring record of the way scholars think about their field, the way they conduct research, the way they organize their data, and generally the assumptions implicit in their taxonomies and interpretations. An ambitious integrative project could gather such information from scattered studies as well as summarize the study outcomes to assess the scientific enterprise in a given area.

Glass has endeavored to apply meta-analysis in this way. It is not surprising that Glass's meta-analytic reviews would be taken by other researchers as models of a standard method. After all, he first articulated the evolution of the quantitative review and he applied his considerable analytical talents to questions of broad significance in reviews of his own. The meta-analyses of psychotherapy effectiveness and class size influence are classics of research integration (Smith, Glass, & Miller, 1980; Glass, Cahen, Smith, & Zelby, 1982). Other reviewers have followed his lead (e.g., Hartley, 1977/1978; Hansford & Hattie, 1982; White, 1982).

Glass assumed that researchers had proven more than they know, that they had already gathered enough data to definitively answer questions of considerable consequence. Researchers had not recognized their accomplishment because their work was scattered in studies too limited and too numerous to interpret. By applying quantitative methods to the integration of this work, Glass sought to identify the regularity hidden in the contradictions, to package large amounts of information in an

easily accessible form, and to remove the subjective biases that may taint informal integrative interpretations.

Criticisms of this form of meta-analysis often focus on the liberal standards for study inclusion. Glass has consistently argued that studies should be included in meta-analysis regardless of methodological quality. This flies in the face of integrative tradition where the reviewer will often give greater weight to findings from carefully designed studies than findings from more poorly controlled investigations. Critics claim that the admission of such data contaminates meta-analytic conclusions with invalid research.

Glass's defense is empirical. By pooling data from research of varying quality, the meta-analyst can empirically determine the presence and even the degree of influences produced by methodological flaws. Disregarding studies *a priori* from review because of their quality introduces the biases of the reviewer, removes a large sample of findings which may in fact differ little from more controlled research, and militates against the possibility of discovering some potentially important relationships between methodological variations and study outcomes.

Critics also complain that Glass's form of meta-analysis obscures critical differences among subgroups of studies. For example, in determining an overall measure for psychotherapy effectiveness, Glass combined outcomes from psychoanalytic, behavioral, rational-emotive, and humanistic applications. Critics charge that such mixes of "apples and oranges" are ultimately meaningless.

In response, Glass suggests the inclusion of studies in a meta-analysis should be determined by the purposes of the reviewer. The

meta-analyst may narrowly define his or her inquiry and include only "apples." A general survey of fruit, though, would necessitate mixing various kinds. As Wortman (1983) put it, "If a policy decision to reimburse clinical psychologists for psychotherapy is being considered, than a general question concerning the effectiveness of all therapies seems appropriate" (p. 240). Besides, the average effectiveness measure across treatments can always later be broken down into specific treatment outcome measures.

After all the studies are collected, each study must be represented to the review. Study outcomes must be transformed to a common metric to allow comparisons across the different scales of different dependent measures. Cohen's *d*, the difference between experimental and control means divided by the pooled standard deviation, would make a logical choice. Glass is concerned, however, that since an experimental treatment may increase variability and that variance in the experimental group may be greater than variance in the control group. Assuming the control group's variation is a better estimate of variation in the untreated population, Glass's effect size (*ES*) divides the difference between experimental and control means by the control group's standard deviation.

The study finding is the unit of analysis in Glass's method. That is, one study may produce several findings from several dependent measures. Each finding is transformed into a separate effect size. The more dependent measures a study has, the more frequently it is represented in the overall meta-analysis.

The meta-analyst has defined an area to review, collected all relevant studies, and calculated effect sizes for every study finding.

As the meta-analyst begins to combine findings from different studies, a second kind of "apples-and-oranges" problem arises. Different dependent measures may identify changes in very different constructs. For example, one study may measure the influence of psychotherapy in terms of galvanic skin response, while another may use a patient's subjective report of well-being. Both study the effects of psychotherapy, but they study its effect on different phenomena. Averaging effect sizes across such diversity produces an average that is difficult to interpret.

Glass's reply is pragmatic. The purpose of the meta-analyst should determine how the effect sizes are grouped. Averaging across constructs is defensible if the meta-analyst seeks an overall impression of psychotherapy effectiveness in available research. If the meta-analyst wishes to determine how one type of psychotherapy differentially affects different phenomena, effect sizes could be separately analyzed according to the constructs they represent.

The three products of Glassian meta-analysis are an average of the effects of treatment on all dependent measures, the standard deviation of those effects, and an examination of significant relationships between independent variables and dependent measures. The first two products are simple to calculate; the third is more difficult. Every study is coded on a number of quantitative dimensions (e.g., number of sessions of psychotherapy) and categorical variables (e.g., random assignment of subjects vs. nonrandom assignment). For example, Smith and Glass (1977) coded at least 16 independent variables for each study. Variables are preselected by the reviewer as potential determinants of variability in study findings. They represent differences in treatment application, methodological differences and controls for validity,

differences in publication history, and ecological differences, i.e., differences in setting, subjects, researcher, and so on. Parametric tests are applied to identify relationships between study outcomes and the coded study features.

Glass has been disappointed by the dearth of variables found to significantly explain diversity in study findings (Glass, Smith, & McGaw, 1981). However, Glass has more recently restated his confidence in the general robustness and usefulness of his method (Glass & Kleigl, 1983). He also made his view of the purposes of meta-analysis explicit. The goal of meta-analysis is larger than the mere estimation of the effectiveness of a treatment or summarizing a sample of research according to Glass. It endeavors "to determine how and in what ways the judgements, decisions, and inclinations of persons (scholars, citizens, officials, administrators, policy makers) ought to be influenced by the literature of empirical research" regarding a defined enterprise (p. 35). To accomplish this, the meta-analyst evaluates, not only a treatment and its effect, but the method of research and the taxonomical structure employed by those engaged in work in this field. In this way, meta-analysis takes on a more encompassing task.

The Meta-analytic Literature Review

In contrast, a meta-analysis might have more restricted goals. That is, a reviewer may wish to integrate scattered studies specifically to describe the most accurate representation of these studies' findings. This is, of course, not inconsistent with Glass's purposes, simply more focused. J. Kulik is perhaps the most prolific advocate of this meta-analytic type.

Kulik and his colleagues have produced about 15 meta-analyses since 1979. Most of these evaluate the success of various instructional technologies (e.g., J. Kulik, Kulik, & Cohen 1979; Kulik & Bangert-Drowns, 1983-1984), though some examined the effectiveness of test coaching (e.g., Bangert-Drowns, Kulik, & Kulik, 1983). Kulik has never published a theoretical exposition of his method, but a perusal of his work shows that his is a consistent and coherent alternative to Glass's approach.

Methodologically, Kulik's approach might be called a more conservative variation on Glass's theme. Kulik is more concerned than Glass about the "apples-and-oranges" problem. Rather than choose a generic topic to investigate, such as psychotherapy, Kulik's foci are typically more specific, such as the effectiveness of the Personalized System of Instruction (J. Kulik, Kulik, & Cohen, 1979) or achievement outcomes related to uses of computer-based instruction in secondary schools (Kulik, Bangert, & Williams, 1983).

More importantly, Kulik is concerned with the "apples-and-oranges" problem at the level of effect size combination. If dependent measures appear to gauge different constructs, Kulik and his colleagues will analyze them separately. For example, if a collection of studies measure students' achievement test scores, attitudes toward computer-based instruction, and attitudes toward course content, analyses of effect sizes for each type of dependent measure will be kept separate.

Kulik is more exclusive than Glass in selecting studies for meta-analytic review. If any study suffers from flaws that may obscure treatment effects or if the study's method deviates too greatly from standard experimental practice, the study is rejected. All studies must

possess experimental and control groups. If subjects are not randomly assigned to these groups, there must at least be evidence that the two groups show pretreatment equivalence. Any study whose treatment "teaches the test," that is, specifically trains the experimental group on the dependent measure, is excluded. The experimental treatment in each accepted study must roughly fit criteria established by the recognized experts in the area under review. Control groups must undergo treatment that is roughly equivalent across studies. With Kulik's work, this usually means that all control subjects are given what would be considered conventional teaching by most educators.

Like Glass, Kulik uses quantitative and quasi-quantitative variables to code study features as independent variables. Glass's *ES* is used as an outcome measure and parametric statistical tests are used to identify relationships between study features and study outcomes. Glass and Kulik do differ, however, in the unit of analysis.

The unit of analysis is the study in Kulik's method. This ensures the statistical independence of each effect size. Effect sizes will not be independent if more than one can come from one study, as in Glass's method. In addition, by representing a study only once to each analysis, no single study carries more weight than any other. When Glass uses the study finding as the unit of analysis, he gives greater weight to studies with more dependent measures, producing an arbitrary bias.

A number of other researchers have expressed goals consistent with Kulik's. In an early formulation, Mansfield and Busse (1977) presented a four-step approach to meta-analysis as an alternative to Glass's method. Their alternative was designed specifically to avoid

inappropriate aggregations of "apples and oranges" and to protect meta-analytic results from contamination by studies with severe methodological flaws. Severely flawed studies were to be eliminated from the analysis; other studies were to be divided into groups of well-designed and methodologically limited research and tested for differences.

Wortman and Bryant (in press) reflect the same concerns as Kulik in their meta-analysis of school desegregation studies. In their review, the authors represented each study once in the overall effect size. They selected studies whose outcomes were measured on the same type of dependent measure, school achievement tests. Even more than Kulik, these authors developed thorough and explicit criteria for study exclusion. The 33 criteria can be broadly categorized as threats to construct validity, external validity, internal validity, and statistical conclusion validity. Their purpose was to reasonably describe the effects of school desegregation on black students' achievement, as shown in available literature.

Landman and Dawes (1982) used Kulik's approach to meta-analysis to reanalyze Smith and Glass's (1977) psychotherapy review. They acquired a random sample of the studies used by Smith and Glass and eliminated from the sample those studies with inadequate control. They compared results obtained when using the finding as the unit of analysis to results based on the study as the unit of analysis. While study quality and nonindependence of data had little effect on the overall findings, Glass's averaging across different kinds of independent and dependent measures appears to have obscured some potentially important relationships.

The differences between Kulik's method and Glass's method are few, but they are significant. They betray the different purposes in these meta-analytic approaches. Glass, with his liberal standards of inclusion and empirical attitude toward methodological flaws and differences in dependent criteria, seeks to use quantitative means to assess a scholarly subculture, to examine modes of inquiry and construct definition as well as estimate treatment effects. Kulik, on the other hand, seeks to produce a reasonable representation of a treatment's effectiveness in the context of previous research. His focus is narrower than Glass's, but he also avoids a potentially hubric effort, the attempt to estimate population effects.

Kulik's method is very consciously a literature review, a quantitative summary of the conclusions of colleagues. The quality, quantity, and historical context of the research reviewed imposes limitations on any estimation of population treatment effect and Kulik seems especially sensitive to these limits. Describing his meta-analytic work on computer-assisted instruction, Kulik put it, "Although these findings are impressive, they are of the past, not necessarily of the future... Meta-analyses lag especially far behind the leading edge in development. They provide at best a summary of the major themes in reported evaluation findings" (Kulik & Bangert-Drowns, 1983-1984, p. 155).

The Combined Probability Method

In 1976, Rosenthal published his book Experimental effects in behavioral research. As an appendix to the book, Rosenthal presented a quantitative synthesis of over 300 studies in the area. Although he appears to have developed his method independently of Glass, his

approach can in retrospect be identified as a primitive form of meta-analyses.

Rosenthal (1976) employed two measures in his survey. First, he cumulated the probability estimates of the outcomes of all the studies. The cumulated probability indicates the likelihood that differences found in the cumulated pool of subjects occurred by chance. Secondly, Rosenthal averaged estimates of effect size obtained from each study to assess the magnitude of the treatment's influence.

Since 1976, Rosenthal and his colleagues (Rosenthal, 1978, 1983; Rosenthal & Rubin, 1982b, 1982c; Cooper, 1979) have continued to develop answers to meta-analytic methodological concerns. They have in fact developed two distinguishable meta-analytic methods; the more recent form is discussed later in this paper. The combined probability method, discussed here, has already been used to survey a number of areas (e.g., Arkin, Cooper, & Kolditz, 1980; Cooper, 1979; Dusek & Joseph, 1983; Rosenthal, 1976; Rosenthal & Rubin, 1978; Wolf, Savickas, Saltzman, & Walker, 1983). Since it is very simple to use, features of this approach can be employed in narrative review as well as meta-analysis.

Rosenthal (1978) investigated nine different procedures for probability combination. Some of these cumulation techniques are fairly primitive. One method involves performing a sign test on probability values divided into two groups, those above .50(+) and those below .50(-). Other methods test the probability of the chance occurrence of the mean of the *p*-values or the mean of *Z*s corresponding to each study's *p*-value. Two other methods based on the summation of the original *p*-values or of their logarithmic transformations are limited to only small groups of studies.

Rosenthal (1978) identified the Stouffer method as simple to calculate and "most serviceable under the largest range of conditions" (p. 191). To use this method, the reviewer must obtain an exact one-tailed probability value for the outcome of each study, then find the corresponding standard normal deviate. These Zs are added, then divided by the square root of the number of studies being combined. The resulting cumulative Z defines the probability that the pooled subjects would be distributed among treatments as they are in the collected studies. The greatest difficulty with this method is that often primary researchers do not publish sufficient data to permit the calculation of exact p-values.

Rosenthal has extended the Stouffer method to address what he calls "the file drawer problem" (Rosenthal, 1978, 1979). A reviewer may be concerned that the outcomes of some unlocated group of studies may change his or her meta-analytic conclusions. Rosenthal's "fail-safe N" indicates the number of additional studies of no effect needed to reduce a significant combined probability to nonsignificance.

The formula for calculating the fail-safe N is an algebraic manipulation of Stouffer's formula. According to Stouffer's formula, the cumulative Z is equal to the sum of the Zs for each study divided by the square root of the number of studies:

$$Z_{cum} = \frac{\Sigma Z}{\sqrt{N}}$$

Set Z_{cum} at 1.645, the Z associated with $p < .05$, one-tailed. The reviewer can calculate ΣZ from the probabilities associated with each study. Solving for N, one can calculate the maximal number of studies one could collect and still have $Z_{cum} = 1.645$.

$$1.645 = \frac{\Sigma Z}{\sqrt{N_{max}}}$$

$$N_{max} = \left[\frac{\Sigma Z}{1.645}\right]^2$$

$$N_{max} = N_{located} + N_{fail\text{-}safe}$$

N_{max} is the maximal number of studies whose sum of Zs, ΣZ, would still produce a significant Z_{cum}. Next, the actual number of studies used to calculate ΣZ is subtracted from N_{max}. The remainder identifies the number of unlocated studies with no effect (that is, studies which would contribute nothing to ΣZ) that could be added to the meta-analysis without changing a significant overall probability to a nonsignificant one. The reviewer is warned that the fail-safe N represents the number of unlocated studies of *zero effect*. There may, in fact, be a smaller number of "*negative*" studies (studies favoring the control group) that would reduce the meta-analytic outcome to nonsignificance.

Meta-analyses based on these techniques have a typical format (e.g., Cooper 1979). The reviewer specifies an area of interest and all relevant studies are collected. For each study, exact p-values are calculated and their corresponding standard normal deviate (Z) listed. Additionally, Cohen's d is calculated for each study as a measure of effect that is more independent of sample size than is a p-value. Effect sizes are derived directly from means and variances when given, otherwise from t or F statistics. (Cohen's d is sometimes also represented as Cohen's U_3, the percentage of control group subjects with dependent measures less than the mean of the experimental group.) Finally, summary measures are taken. Stouffer's method is used to

determine the overall probability of this distribution of subjects and the fail-safe N is calculated. The ds (and U_3s) are averaged to provide an overall effect size measure.

The averaged effect size and a combined probability measure are the outcomes of this type of meta-analysis. The averaged d offers an estimate of the treatment effect found in the studies. Stouffer's Z, representing the combined interactions of effect size and sample size from all the studies, is an indicant of the reliability of these findings. Cohen's U_3 and the fail-safe N are more intuitively understandable forms of the overall d and Z.

The unit of analysis in the calculation of the average d is clearly the study. Each study contributes one effect size and these are simply averaged across studies. Stouffer's Z, however, is not a simple average across studies. It is, in fact, a standard normal deviate, locating the cumulative Z in a distribution of cumulative Zs. The distribution is assumed to have a mean of zero. Its variance is equal to the sum of the variances of each original Z_j. Since each Z_j has unit variance, the sum of variances is simply the number of studies, N.

$$\text{Stouffer's } Z = \frac{\Sigma Z_j - 0}{\sqrt{\text{Var. } \Sigma Z_j}} = \frac{\Sigma Z_j}{\sqrt{\Sigma \text{ Var. } Z_j}} = \frac{\Sigma Z_j}{\sqrt{N}}$$

since Var. $Z_j = 1$.

What is the unit of analysis of Stouffer's Z? The Z corresponding to each study is influenced by effect size and sample size. That is, the higher the effect size, the lower the p-value and the higher the corresponding Z. However, the effect sizes of a group of studies being combined are assumed to be roughly equal. Sample size varies. The

larger the sample size, the smaller the *p*-value and the higher the corresponding *Z*. *Z* is therefore a subject-sensitive statistic, and the subject is ultimately the unit of analysis for Stouffer's *Z*.

The *Z*s can be made even more sensitive to the number of subjects if each *Z* is weighted by the proportion of the study's *N* to the overall *N* (Rosenthal, 1978; Cooper, 1979). This may not be a reasonable weighting, however, in most applications. Assuming that the effect sizes of the studies are roughly equivalent, *Z*s will only differ according to differences in sample size. A *Z* that is already large due to a large *N* will only be exaggerated if that *N* is also used as a weight.

What is most conspicuously absent from this form of meta-analysis is rigorous attention to differences in study features. This method is not as interested in the regularities of outcome variation as it is in the degree to which outcomes converge. Decisions regarding the inclusion or exclusion of methodologically inferior studies is left to the reviewer's discretion. Only very basic differentiations are made in a literature, for example, differentiating between studies examining race and studies examining sex as bases of teacher expectation (Dusek & Joseph, 1983). Other differentiations may be made (e.g., Rosenthal [1976] compares dissertation and published research and well-controlled and less controlled studies), but these are typically done to determine the most accurate estimate of an overall treatment effect.

Strictly speaking, this method is similar to Glass's and Kulik's in that its conclusions are bound to the studies it collects. Effect size is averaged from a specific group of studies; the probability measure is

based on the subjects pooled from the same specified group. In this sense, Rosenthal's method is a literature review.

However, the method intends to do more. The implicit purpose of Rosenthal's method is not to review a literature but to estimate a reliable treatment effect and provide some measure of its reliability. These measures are intended to be broadly generalizable. Cooper (1979), for example, went so far as to generalize his meta-analytic findings to predictions of the results of future research. In keeping with this goal of deriving generalizable conclusions, Rosenthal's method has begun to focus on the subject as the unit of analysis, an effort to draw conclusions from large samples of subjects. Rosenthal's combined probability method can be seen as a transitional form of meta-analysis between, on the one hand, meta-analyses whose conclusions are interpreted as specific to the studies reviewed and, on the other hand, meta-analyses that attempt to determine population parameters and "true" treatment effects.

The Analysis of Study Variation Using Tests of Homogeneity

Rosenthal and Rubin and Hedges appear to have developed another form of meta-analysis independently. Though there are some differences between them, there is general agreement in purpose and statistical technique. Their applications of homogeneity tests to meta-analysis are considered together here.

Kulik has modified Glass's method for his own purposes. Hedges, on the other hand, sees his method as a correction of Glass's. Glass's "methods may be reasonable *ad hoc* procedures, but are generally not appropriate" (Hedges & Olkin, 1982, p. 163). Hedges's discomfort with Glass's analyses may partly be due to a misreading of Glass's purposes.

Hedges assumes that Glass wishes to estimate "true" population treatment effects. As a method of such estimation, Glass's method "does not go far enough" (Hedges & Olkin, 1982, p. 162). Hedges therefore proposes some alternatives.

First, Hedges identified Glass's effect size measure (as well as Cohen's d) as a biased estimator. Glass's effect size, designated "g," has a noncentral t-distribution and is therefore asymmetric, nonnormal, and positively skewed when the population effect is not zero. This bias is most severe when g is based on a small sample. Hedges has formulated a correction factor to eliminate this bias. Multiplying g by this correction factor produces an unbiased estimator of the population treatment effect, g^U. Not surprisingly, the distribution of g^U has less variance than the distribution of g.

The derivation of g^U may have some theoretical importance but it is of dubious practical significance. As Hedges (1982c) puts it, "It is clear that as m [degrees of freedom] becomes large, g^U tends to g, so g is almost unbiased in large samples" (p. 492). Rosenthal and Rubin (1982b) comment, "As long as each ν_j [degrees of freedom] is larger than, say, 10, the bias...is of little consequence, being less than 8%; furthermore, if all the studies have approximately the same ν_j, the bias...will be the same for all studies under the hypothesis of constant effect size. Hence, d_j is effectively unbiased" (p. 504).

Bangert-Drowns, Kulik, and Kulik (1983) calculated g and g^U for 27 studies on the effectiveness of achievement test coaching. They found a correlation of .999 between the "biased" and "unbiased" estimators. Hedges and Stock (1983) reanalyzed data from Glass's meta-analysis of

the effect of class size. Using the unbiased estimator, however, produced no significant changes in Glass's original assessment.

Hedges (1981) has also recommended corrections to effect size measures for the unreliability and invalidity of dependent measures. Unreliability will always reduce an effect size and is easily corrected if the dependent measure's reliability is known. The effects of invalidity are not so easy to characterize, and they can only be corrected if a valid response measure of known reliability has a known correlation with the dependent measure. Needless to say, the meta-analyst is rarely given enough information to correct the effects of either unreliability or invalidity.

One correction to the effect size that both Rosenthal and Hedges recommend is the use of weighting. A reviewer can give larger weights to effect sizes that are based on larger samples and are thus less subject to sampling error. This would increase the accuracy of the cumulated effect size. The weights that would most minimize the variance of the cumulated effect size depend on the sampling variances of the individual effect sizes, which in turn depend on the sample sizes of each experiment and the population effect sizes (Hedges, 1981). The ideal weight for each effect size is the proportion of the inverse of the effect size's sampling variance to the sum of the inverses of all the effect sizes' sampling variance (Hedges, 1981; Rosenthal & Rubin, 1982b, 1982c). This is the same weight discussed by Cochran (1937) and Yates and Cochran (1938) for integrating agricultural research. This weight makes estimation more accurate only in situations where the "real" treatment effect is the same in each study.

$$W_i = \frac{\frac{1}{\text{Var. } (d_i)}}{\Sigma\left[\frac{1}{\text{Var. } (d_i)}\right]}$$

Hedges (1981) has shown that when samples are large and experimental and control groups equal in size, this weight is the simple proportion of an effect size's degrees of freedom to the sum of the degrees of freedom of all the effects sizes. Hunter, Schmidt, and Jackson (1983) are satisfied that the frequency-weighted effect size, that is, the effect size weighted by the sample size divided by the sum of all the sample sizes, is an adequate approximation to the optimally weighted effect size.

Glass's approach to the study of variability in effect sizes is to use familiar statistical tests, *t*-tests, ANOVA, linear regression, and the like, to identify reliable patterns in study outcomes. This approach is not appropriate to Hedges's and Rosenthal's purpose. Glass's approach is designed to assess summary statistics in a literature review, not identify population treatment effects. Secondly, Glass's approach is meant to test differences between *a priori* categories but provides no overall test to determine if such categorical tests are necessary at all. Third, Glass's approach focuses on variability among studies, not on differences in within-study variability. If a meta-analyst seeks to use individual effect sizes as estimates of treatment parameters, then differences in their variability, particularly in the variability due to sampling error, indicate that they differ in reliability as estimators. Such differences should be given attention.

Hedges (1982c) and Rosenthal and Rubin (1982b) have devised a test consistent with their purposes. Variability among study outcomes is represented as the sum of squared differences between each effect size and the weighted average effect size. The ratio of this sum of squared differences to the sum of sampling variance in all the effect sizes is distributed as a chi-square. If this ratio, designated "H", is approximately 1, variation among studies is approximately the same as sampling error within studies. If H is large or statistically significant, the variation among studies cannot be attributed to sampling error only.

Hedges (1982a) extends the use of this test of homogeneity. If the overall H is large, the effect sizes can be broken into previously determined categories. H can then be calculated for each category. If H within a category is small, the effect sizes within the category are assumed to be homogeneous and represent a single population treatment effect. A variant of the H statistic, H_B, is given to represent the degree to which the categories vary among each other. The overall H, within-class H, and between-group H_B are analogous to the partitions of variation in ANOVA. The overall H is equal to the sum of the within-class H and between-group H_B.

To test for significant differences among effect sizes of a heterogeneous group, Rosenthal and Rubin (1982b, 1982c) suggest a simple formula for contrasts among effect sizes. The concept is analogous to contrasts in analysis of variance. The sum of appropriately weighted effect sizes is distributed as a standard normal deviate. This Z can be easily tested for statistical significance.

The *N* statistic can be reformulated to apply to continuous data. Hedges (1982b) shows how beta-weights can be estimated for a variant of linear regression that is sensitive to sampling error in the effect sizes used as data points. Statistics analogous to the weighted sum of squares due to regression and to the weighted residual sum of squares can then be calculated. These statistics are distributed as chi-squares and can be tested for significance.

The use of the test of homogeneity bears some resemblance to Light and Smith's (1971) cluster analysis. The meta-analyst using this method goes through a cycle of testing for homogeneity, partitioning the data, and testing again. Cluster analysis similarly attempted to group data into natural clusters of homogeneity, only cluster analysis is a secondary analysis of original data. Hedges and Rosenthal have weighted effect size to reflect differences in sample size and sampling error. These subject-sensitive statistics are clustered and reclustered as they form natural aggregations of greater homogeneity, tested by subject-sensitive significance criteria. The subject is the unit of analysis even though this unit is mediated by the summary statistic, effect size.

Hedges and Rosenthal have concentrated primarily on statistical methodology. However, the method they advocate has certain implications for nonstatistical procedural issues. Hedges, for example, appears to be as willing as Glass to include studies of diverse quality and treatment features. The tests of homogeneity can be used to empirically determine if this diversity produces heterogeneous outcomes (Hedges & Olkin, 1982).

However, Hedges, like Kulik, will not allow the averaging of effects from dependent measures representing different constructs.

Hedges and Rosenthal are attempting to estimate single population effects and therefore effect size estimates must come from dependent measures that are roughly linearly equitable. "Even if a treatment produces a uniform effect size on measures on one construct (such as mathematics achievement), there is little reason to expect that effect size to be the same as effect size for studies that measure the influence of the same treatment on another construct (such as attitude)" (Hedges, 1982c, p. 491).

The practical usefulness of tests of homogeneity is unclear. Rosenthal and Rubin (1982b) warn that the behavior of statistics used in this meta-analytic type is not fully understood for small or nonnormal samples. Hedges and Stock (1983) reexamined the continuous model offered by Glass relating class size to student achievement. Although they used modified beta-weights and modified estimates of residual sum of squares and sum of squares due to regression, their conclusions largely duplicated the earlier findings of Glass.

In addition, some authors have spoken against the use of tests of homogeneity. Hunter, Schmidt, and Jackson (1983), for example, developed a test of homogeneity for use with the correlation coefficient. This test is identical to Hedges's and Rosenthal's. However, Hunter, Schmidt, and Jackson (1983) have commented, "We do not endorse this significance test because it asks the wrong question. Significant variation may be trivial in magnitude, and even nontrivial variation may still be due to research artifacts...it has very high statistical power and will therefore reject the null hypothesis given a trivial amount of variation across studies" (p. 46).

The Analysis of Study Variation Using Estimates of Population Variation

In many ways this fifth and last form of meta-analysis is a variation of the one advocated by Rosenthal and Rubin and Hedges. The two methods agree in basic purpose and analytic strategy. Their greatest difference is that Hunter and Schmidt will not use tests of homogeneity. However, this type of meta-analysis deserves separate attention. It possesses a kind of flexibility and straightforwardness that may make it increasingly appealing and versatile for reviewers in the future.

Hunter and Schmidt became interested in quantitative research synthesis as a tool to investigate the differential validity of employment tests for blacks and whites (Schmidt, Berner, & Hunter, 1973; Hunter & Schmidt, 1978; J. Hunter, Schmidt, & Hunter, 1979). Studies in this area reported inconsistent findings, some indicating that employment tests were valid regardless of race, others indicating differences in validity. In the course of their review, it became apparent that most of the variation among the correlations was due to research artifacts. "On the average across test-job combinations, we find that 72% of the variation across studies is accounted for by four artifacts: sampling error, variation in criterion reliability, variation in test reliability, and variation in range restriction." The remaining variation, they believed, could be attributed to reporting errors. "Our results show that the previously widespread belief that cognitive employment tests are frequently invalid is a myth created mostly by sampling error" (Hunter & Schmidt, 1981, p. 7).

Not surprisingly, features of the Hunter-Schmidt method of meta-analysis reflect this earlier history. For example, these authors

suggest that the correlation coefficient (r) is a superior measure of effect since it can so easily measure "the variance accounted for" (r^2) and can be inserted into correlation matrices for larger analyses. Their earliest meta-analytic strategies were therefore designed for use with the point-biserial correlation. Similarly, it is not surprising that the goal of their strategy would be to correct study outcomes for research artifacts to produce the best estimates of population treatment effects.

Since d is already a widely used statistic in meta-analysis, easily transformed into r, and well suited to two-group experimental comparison, Hunter and Schmidt's most recent formulations make more use of d. Their meta-analytic procedures include many familiar stages. They are most comprehensively present in Hunter, Schmidt, and Jackson (1983).

First, studies are collected, and effect size is calculated for each study. All studies that bear on the question of interest are used regardless of any criteria of methodological adequacy. Hunter and Schmidt argue that all studies have methodological inadequacies and that the meta-analyst can empirically determine which inadequacies produce biased effects only if they are included in review. One effect size is used to represent each study unless the purposes of the meta-analyst specifically warrant the separation of effect sizes for different dependent measures.

All effect sizes are individually corrected for unreliability or other statistical artifacts when appropriate information is given for each study. Such information is not always available. A meta-analyst assessing the effectiveness of individualized instruction in the

secondary schools, for example, might collect 50 studies, but only 20 may report reliabilities. Hunter and Schmidt recommend using the average of the reliabilities found in the 20 studies to correct all the effect sizes.

Second, the variation of the effect sizes is studied. The frequency weighted sum of squared difference between each effect size and the frequency weighted average effect size measures the overall variability of the outcomes.

$$\sigma_d^2 = \frac{\Sigma[N_i(d_i - \bar{d})^2]}{\Sigma N_i} \quad \text{where } \bar{d} = \frac{\Sigma N_i d_i}{\Sigma N_i}$$

The amount of variation attributable to sampling error is calculated next. This calculation is identical to the sampling error approximations used by Hedges and Rosenthal. Here, however, Hunter and Schmidt do not test for homogeneity. Instead, they subtract the variation due to sampling error from the total variation. If sampling error accounts for as much as 75% of the overall variation, the effect sizes are assumed to estimate one parameter (i.e., they are homogeneous), and the frequency weighted effect is the best estimate of the parameter.

If the variation is still greater than zero after correction for sampling error, the meta-analyst should investigate the influence of moderator variables. The meta-analyst develops and codes a set of independent variables representing study and treatment features suspected of producing outcome variation. Effect sizes can be divided into groups according to these categories. If a variable really acts as a moderator, the means of the categories should be different, and their

variation should be less than it was when they were mixed. Alternatively, correlations can be taken between effect sizes and study features (and these correlations can be corrected for sampling error in the effect sizes) to assess the influence of moderators.

This type of meta-analysis has a number of products. There is an overall mean and standard deviation corrected for statistical artifacts. Once moderating variables have been identified, Hunter and Schmidt recommend the use of regression on the effect sizes as individual data points in much the way Glass recommends. Such a procedure allows the meta-analyst to estimate the distribution of the treatment effects when study characteristics are at their means and when they are varied.

The Hunter-Schmidt procedure is consistent with the goals of meta-analysis described by Rosenthal and Hedges. These four researchers wish to base causal observations on larger groups of subjects. When primary data is not available, meta-analysis can approximate cluster analysis by grouping effect sizes into natural aggregations. However, instead of using tests for homogeneity of statistical significance, Hunter and Schmidt resort to real measures of variance and approximate sampling error. Using these, the meta-analyst is more aware of the magnitude of variation being examined, not just its statistical significance.

Furthermore, use of the real measures allows for more creative manipulations of study variation. DerSimonian and Laird (1983), for example, used a meta-analytic procedure which, though developed independently, followed the Hunter-Schmidt method fairly closely. They showed that in studies of Scholastic Aptitude Test (SAT) coaching, sampling error could be estimated more exactly than the estimation routinely used in the test of homogeneity. They estimated sampling

error from gain scores, that is, with variation reduced by information about subjects' pretreatment aptitude. Thus, by not applying statistical tests, DerSimonian and Laird more flexibly and accurately examined the magnitude of variation produced by different categories of effect sizes.

Discussion

Research outcomes vary in ways that make generalizable interpretations difficult. Such variation comes from a number of sources. It may reflect real population variation, the effects of different treatment features or study settings, sampling error, selection biases of the reviewer, publication biases, the effects of erroneous or insufficient reporting (unreported spurious influences, computational errors, typographical errors), differing degrees of validity and reliability in the outcome measures, and differences in the range or intensity of the independent variable. The task is enormous, but the power of social scientific inquiry would greatly increase if general principles could be found amid this outcome variation.

A number of processes have been developed to attempt this task. These processes can be arranged on a continuum according to the degree of specificity in quantification. On one extreme, there is the informal narrative review which, in its simplest form, lists the outcomes of some sample of studies and presents the reviewer's reasoned speculations regarding their interpretation. On the other extreme, there are the forms of secondary analysis, such as cluster analysis (Light & Smith, 1971). Cluster analysis pools the original data found in any group of studies bearing on a particular question into smaller, natural

aggregations. Where no differences are found among aggregations, original data is pooled into samples of larger sizes.

Meta-analysis lies between these two extremes. Its procedures are more approximate than those of cluster analysis. Yet, meta-analysis is more statistically rigorous than is the typical narrative review. It attempts to identify some central tendency in the outcomes of a group of studies and additionally analyze their variation without having to resort to original data.

Since it is a new approach to research integration, meta-analysis is still undergoing definitional refinement. There is, however, already a fair degree of agreement regarding its general method. A reviewer gathers all of the studies bearing on a certain topic. A reviewer may exclude studies on any grounds that are consistent with her or his purposes and explicitly stated. The outcomes of the studies are represented by a measure that is independent of the different dependent criteria. These outcomes are averaged to measure central tendency. Typically the meta-analyst will go further and look for relationships between the study outcomes and study features as an explanation for study outcome variability.

While this consensus exists, different meta-analytic methods can be distinguished, at least five of them. These different approaches can be arranged on the same continuum as the other forms of research integration. On the one hand, there are those meta-analyses whose purpose is primarily to review a body of research. These lie more closely to the narrative review. These meta-analysts seek to answer questions of the form "What does available research say about treatment

X's effectiveness?" (Glass, McGaw, & Smith, 1981; J. Kulik, Kulik, & Cohen, 1979).

On the other hand, there are those meta-analyses whose purpose is primarily to increase the sample size used to test a specific hypothesis and to determine a generalizable es imate of a treatment effect. These lie more closely to cluster analysis. While strictly speaking they can only summarize what the research indicates, these meta-analysts seek to answer questions of the form "What is the real effect of treatment X?" (Hedges, 1980; Rosenthal & Rubin, 1982b; and Hunter, Schmidt, & Jackson, 1983).

This difference in purpose is real and dramatic. Glass, representing the "literature review" type of meta-analysis, has pointedly distinguished between the purposes of primary researchers and meta-analysts. Primary researchers seek to develop theoretical explanations for the effects of treatments on subjects. Meta-analysts work to summarize those effects and assess the methods used to test those effects to better inform decision-makers.

> The literature on psychotherapy outcomes is distinct from psychotherapy outcomes themselves by a sequence of translations too obvious to enumerate...The [meta-analytic] inquiry that takes as primary objects the documented outcomes of psychotherapy and the methods of psychotherapy-outcome researchers and that takes as explananda the policy decisions that can be justified by the literature of outcome research will necessarily use taxonomies and methods different from [prima.y research]...We could not have shared many of the concerns that motivated the authors of the primary studies we integrated, their concerns being primarily the

construction of a big theory about changing behavior. (Glass & Kliegl, 1983, p. 35)

On the other extreme, meta-analysts who stand closer to the cluster analysis/data pooling/secondary analysis traditions more closely identify their work with the work of primary researchers. These meta-analysts are attempting to answer the same kinds of questions as the primary researchers, only these meta-analysts hope to use larger samples. "In meta-analysis, the parameter of interest is the population standardized mean difference" (Hedges, 1980, p. 25), "Only cumulation of results across studies can generate [adequate] sample sizes in most areas of contemporary research" (Hunter & Schmidt, 1981, p. 6), and "Meta-analysis...is tailored to situations where only summary statistics are available. Because we have calculated a measure of sampling variations associated with [the study outcomes], we are able to extend traditional limits of meta-analysis to obtain results previously available only when the original data were analyzed" (DerSimonian & Laird, 1983, p. 6). The difference between meta-analysis as a literature review and meta-analysis as an approximation of cluster analysis or data pooling is great.

Even within these two larger subtypes, greater distinctions can be made in meta-analytic method. Glass and Kulik, both espousing methods of literature review, are procedurally in close agreement. We have seen that Kulik's method is consistently more conservative, being more exclusive in the studies chosen for review and more cautious in effect size estimation. The fundamental distinction between these two is not their procedural differences, but their differences in purpose. Glass is most interested in broad questions of the form "What does the

research say about the way social scientists investigate treatment X?" Kulik's focus is more specific: "What does the research say about the effectiveness of treatment X?"

Kulik's and Glass's methods are procedurally similar though their purposes diverge. The opposite can be said about the method developed by Schmidt and Hunter compared to Hedges's and Rosenthal and Rubin's. That is, these latter researchers are generally agreed as to the purpose of meta-analysis: the estimation of the distribution of population treatment effects. However, these researchers diverge on procedural concerns. Hedges (1980) and Rosenthal and Rubin (1982b) advocate the use of a test of homogeneity to partition effect sizes into groups assumed to represent the same parameter. Hunter, Schmidt, and Jackson (1983) advocate the use of the observed variation of the effect sizes and their sampling error. Working with the actual estimates of variability, the meta-analyst is made aware of the magnitude of the differences among the studies and may be able to treat the variation more flexibly.

Rosenthal's combined probability method (Rosenthal, 1976) is a bit more difficult to classify. It is a very primitive form of meta-analysis in that it does little to relate study features to study outcomes and it emphasizes the summary of probability values. It also resembles meta-analysis as a literature review since its overall effect size does not account for differences among study sample sizes. However, the Stouffer Z is related to the number of subjects pooled and conclusions drawn from this type of meta-analysis are meant to be generalized. The combined probability method is probably best understood as a simple form of meta-analysis approximating data pooling.

Each of these methods has its own particular strengths and weaknesses. It might be important to make two general observations though. First, Glass's original procedures are proving to be generally robust. In situations where a body of data is analyzed using Glass's methodology and some other techniques, the differences have consistently proven inconsequential (Bangert-Drowns, Kulik, & Kulik, 1983; Hedges & Stock, 1983). This evidence is too scanty to be conclusive at this point, but it is suggestive.

A second important observation is that all meta-analysts must contend with publication history. It has been repeatedly shown that effect sizes taken from published articles are consistently higher than those found in dissertations (Smith, 1980; Bangert-Drowns, Kulik, & Kulik, 1984). Why this is so is not entirely clear. What is clear is the forbidding possibility that a meta-analyst's conclusion may say more about editorial preference, the politics of research finance, or the differential capabilities and biases of professionals and graduate students than about real treatment effect. This dilemma is more serious for meta-analyses that propose to estimate treatment parameters. Those reviews that more modestly claim only to characterize the available findings can fulfill their purposes whether or not publication histories differentiate those findings.

The overall purpose of meta-analysis is still a topic of debate. There is also continuing debate regarding the most convenient representation of size of effect. This debate operates on three levels. First, some researchers argue for particular effect size displays because they are easily understood or manipulated. Cooper (1979) used Cohen's statistic U_3 in addition to d to represent effect size. U_3

indicates the percentage of control group subjects with outcome scores lower than the mean of the experimental group. Rosenthal and Rubin (1982a) suggest the binomial effect size display (BESD), indicating the change in "success rate" brought about by a treatment. Hunter, Schmidt, and Jackson (1983) suggest that the point-biserial correlation be used as a measure of effect size.

These researchers, and others like them, are right in suggesting that different representations of effect size may be more suitable for particular audiences or research areas. However, in research where experimental and control groups are compared, Cohen's *d* is easily calculated and easy to transform into other representations, such as percentile rank or grade equivalent scores. Cohen's *d* should be taken as the standard measure of effect while other representations can be used as complements for easier reference.

At a second level, there is debate regarding the most accurate representation of effect size. Glass has suggested that the denominator of Cohen's *d* be changed from the pooled standard deviation to the control group standard deviation. He feared that the experimental treatment would increase variation in the pooled standard deviation. Kulik has found, however, a high correlation between Glass's *ES* and Cohen's *d* and has come to use them more or less interchangeably (J. Kulik, Kulik, & Cohen, 1979). Hedges (1980) has argued that *d* must be corrected for bias. However, Bangert-Drowns, Kulik, and Kulik (1983) and Hedges and Stock (1983) have found only trivial differences between *d* and *d* corrected for bias when these are applied to the same data.

Hedges (1980) and Hunter, Schmidt, and Jackson (1983) have advocated other corrections for effect size, including corrections for

unreliability and invalidity of dependent measures and corrections for range restriction. These corrections promise to make real improvements in the accuracy of effect size measurement. However, the information needed to make these corrections is typically unavailable. Again, Cohen's *d* proves a robust and practical estimator of effect.

Thirdly, there is debate over how effect sizes should be cumulated. Glass and Kulik use unweighted averages of effect size. They warn that weighting studies by sample size may give too great a weight to studies that may not be necessarily the most methodologically sound or theoretically representative. Rosenthal and Rubin, Hedges, and Hunter and Schmidt use weighted means since they wish to calculate effect size as if all the subjects were pooled in one large experiment.

There are interesting alternatives to these typical cumulative techniques. Traditional techniques generally assume that effect size is normally distributed. Kraemer and Andrews (1982) developed a nonparametric method to estimate the effect size for each study. The averaged nonparametric effect size would be similar to the simple average or weighted average when subjects are normally distributed but become increasingly different with nonnormal data. Bond and Titus (1983) have used a "five-point presentation" of overall effect size. That is, the maximal and minimal effects as well as the first, second, and third quartile effects are reported to describe effect size distributions. Such a representation is useful if the reviewer suspects the distribution of effects is not normal.

The meta-analyst must choose from these alternatives the most reasonable method for cumulation. There has been virtually no empirical work to compare them. The choice must be determined by the particular

reviewer's conception of the purpose of meta-analysis and of the distribution of effect sizes in the area being reviewed.

Meta-analysts also differ in the ways they handle study features variation. Glass and Kulik enumerate a list of independent variables before beginning a meta-analysis. These study features are coded for each study and they are all examined for statistical relationships to study outcomes. Hedges and Hunter and Schmidt will only examine study features after they have first determined that there is significant variation in outcomes besides variation attributable to sampling error. Only enough features sufficient to explain the remaining variation are examined further.

This difference is a direct product of the more general difference between meta-analysis as a literature review and meta-analysis as a data pooling approximation. Glass and Kulik, reviewing a body of research, treat each study or outcome as a data point, the unit of analysis. Hedges and Hunter and Schmidt, estimating a population treatment effect, treat each effect size not as a data point, but as a summary statistic with its own sampling error. The subject is the unit of analysis here and hence the comparison of the variation among studies to the sampling error in the pooled group of subjects.

These differences in meta-analytic method have gone largely overlooked. It is time they be clarified so that the limitations of this approach to research integration can be more realistically assessed. These differences should not be taken as evidence of some inherent weakness of meta-analysis. It is merely a reflection of the natural evolution of a new social scientific tool. Perhaps meta-

analytic methods appropriate to different purposes and different research areas will ultimately be distinguished.

Conclusion

The meta-analytic perspective has been a familiar part of the social sciences for some time. Meta-analytic method, on the other hand, is relatively new. There is still much debate over general definition and procedure. As with any new method, a long process of rhetoric and empiricism is needed to bring greater clarity to the area.

Meta-analysis is not a fad. It is rooted in the fundamental values of the scientific enterprise: replicability, quantification, causal and correlational analysis. The potential economic losses due to the obscurity of valuable information scattered in individual studies and the inability of social scientists to deliver generalizable answers to basic questions of policy are too great to allow us to treat research integration lightly. The potential benefits of meta-analytic method seem enormous.

References

Arkin, R., Cooper, H., & Kolditz, T. (1980). A statistical review of the literature concerning the self-serving bias in interpersonal influence situations. Journal of Personality, 48, 435-448.

Bangert, R. L., Kulik, J. A., & Kulik, C.-L. C. (1983). Individualized systems of instruction in secondary schools. Review of Educational Research, 53, 143-158.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1983). Effects of coaching programs on achievement test performance. Review of Educational Research, 53, 571-585.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1984, August). The influence of study features on outcomes of educational research. Paper presented at the meeting of the American Psychological Association, Toronto.

Bloom, B. S. (1964). Stability and change in human characteristics. New York: Wiley.

Bond, C. F., Jr., & Titus, L. J. (1983). Social facilitation: A meta-analysis of 241 studies. Psychological Bulletin, 94, 265-292.

Cartwright, D. (1971). Risk taking by individuals and groups: An assessment of research employing choice dilemmas. Journal of Personality and Social Psychology, 20, 361-378.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. Journal of the Royal Statistical Society, Supplement, 4, 102-118.

Cochran, W. G. (1943). The comparison of different scales of measurement for experimental results. Annals of Mathematical Statistics, 14, 205-216.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal and Social Psychology, 65, 145-153.

Cook, T. D., & Leviton, L. C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. Journal of Personality, 48, 449-472.

Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. Journal of Personality and Social Psychology, 37, 131-146.

Cooper, H. M. (1980). The literature review: Elevating its status to scientific inquiry (Tech. Rep. No. 238). Columbia: University of Missouri-Columbia, Center for Research in Social Behavior.

Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. Psychological Bulletin, 87, 442-449.

DerSimonian, R., & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. Harvard Educational Review, 53, 1-15.

Dickson, W. P. (in press). Two decades of referential communication research: A review and meta-analysis. In C. J. Brainerd & M. Pressley (Eds.), Progress in cognitive development research: Vol. 2. Verbal processes in children. New York: Springer-Verlag.

Dusek, J. B., & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. Journal of Educational Psychology, 75, 327-346.

Erlenmeyer-Kimling, L., & Jarvik, L. F. (1963). Genetics and intelligence: A review. Science, 142, 1477-1479.

Fisher, R. A. (1932). Statistical methods for research workers (4th ed.). London: Oliver & Boyd.

Fiske, D. W. (1983). The meta-analytic revolution in outcome research. Journal of Consulting and Clinical Psychology, 51, 65-70.

Glass, G. V. (1976). Primary, secondary, and meta-analysis research. Educational Researcher, 5, 3-8.

Glass, G. V., & Kliegl, R. M. (1983). An apology for research integration in the study of psychotherapy. Journal of Consulting and Clinical Psychology, 51, 28-41.

Glass, G. V., Cahen, L. S., Smith, M. L., & Filby, N. N. (1982). School class size: Research and policy. Beverly Hills, CA: Sage.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Hansford, B. C., & Hattie, J. A. (1982). The relationship between self and achievement/performance measures. Review of Educational Research, 52, 123-142.

Hartley, S. S. (1978). Meta-analysis of the effects of individually paced instruction in mathematics. Dissertation Abstracts International, 38, 4003A. (University Microfilms No. 77-29,926)

Hedges, L. V. (1980). Unbiased estimation of effect size. Evaluation in Education: An International Review Series, 4, 25-27.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6, 107-128.

Hedges, L. V. (1982a). Fitting categorical models to effect sizes from a series of experiments. Journal of Educational Statistics, 7, 119-137.

Hedges, L. V. (1982b). Fitting continuous models to effect size data. Journal of Educational Statistics, 7, 245-270.

Hedges, L. V. (1982c). Estimation of effect size from a series of independent experiments. Psychological Bulletin, 92, 490-499.

Hedges, L. V. (1983). A random effects model for effect sizes. Psychological Bulletin, 93, 388-395.

Hedges, L. V., & Olkin, I. (1982). Analyses, reanalyses, and meta-analysis. Contemporary Education Review, 1, 157-165.

Hedges, L. V., & Stock, W. (1983). The effects of class size: An examination of rival hypotheses. American Educational Research Journal, 20, 63-85.

Hunter, J. E., & Schmidt, F. L. (1978). Differential and single-group validity of employment tests by race: A critical analysis of three recent studies. Journal of Applied Psychology, 63, 1-11.

Hunter, J. E., & Schmidt, F. L. (1981). Cumulating results across studies: Correction for sampling error. A proposed moratorium on the significance test and a critique of current multivariate reporting practices. Unpublished manuscript, Michigan State University, East Lansing.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 86, 721-735.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1983). Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage.

Ide, J. C., Parkerson, J., Haertel, G. D., & Walberg, H. J. (1980). Peer influences. Evaluation in Education, 4, 111-112.

Jackson, G. B. (1980). Methods for integrative reviews. Review of Educational Research, 50, 438-460.

Jones, L. V., & Fiske, D. W. (1953). Models for testing the significance of combined results. The Psychological Bulletin, 50, 375-382.

Kraemer, H. C., & Andrews, G. (1982). A nonparametric technique for meta-analysis effect size calculation. Psychological Bulletin, 91, 404-412.

Kulik, J. A. (1984, April). The uses and misuses of meta-analysis. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Kulik, J. A., Bangert, R. L., & Williams, G. W. (1983). Effects of computer-based teaching on secondary school students. Journal of Educational Psychology, 75, 19-26.

Kulik, J. A., & Bangert-Drowns, R. L. (1983-84). Effectiveness of technology in precollege mathematics and science teaching. Journal of Educational Technology Systems, 12, 137-158.

Kulik, J. A., Kulik, C.-L. C., & Cohen, P. A. (1979). A meta-analysis of outcome studies of Keller's personalized system of instruction. American Psychologist, 34, 307-318.

Landman, J., & Dawes, R. M. (1982). Psychotherapy outcome. American Psychologist, 37, 504-516.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 41, 429-471.

Mansfield, R. S., & Busse, T. V. (1977). Meta-analysis of research: A rejoinder to Glass. Educational Researcher, 6, 3.

Mintz, J. (1983). Integrating research evidence: A commentary on meta-analysis. Journal of Consulting and Clinical Psychology, 51, 71-75.

Mosteller, F. M., & Bush, R. R. (1954). Selected quantitative techniques. In G. Lindzey (Ed.), Handbook of social psychology: Vol. 1. Theory and method. Cambridge, MA: Addison-Wesley.

Pearson, E. S. (1938). The probability integral transformation for testing goodness of fit and combining tests of significance. Biometrika, 30, 134-148.

Pearson, K. (1933). On a method of determining whether a sample size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. Biometrika, 25, 379-410.

Rosenthal, R. (1976). Interpersonal expectancy effects: A follow-up. In R. Rosenthal, Experimental effects in behavioral research (pp. 440-471). New York: Irvington.

Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. Psychological Bulletin, 86, 638-641.

Rosenthal, R. (1983). Assessing the statistical and social importance of the effects of psychotherapy. Journal of Consulting and Clinical Psychology, 51, 4-13.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. The Behavioral and Brain Sciences, 1, 377-386.

Rosenthal, R., & Rubin, D. B. (1982a). A simple, general purpose display of magnitude of experimental effect. Journal of Educational Psychology, 74, 166-169.

Rosenthal, R., & Rubin, D. B. (1982b). Comparing effect sizes of independent studies. Psychological Bulletin, 92, 500-504.

Rosenthal, R., & Rubin, D. B. (1982c). Further meta-analytic procedures for assessing cognitive gender differences. Journal of Educational Psychology, 74, 708-712.

Schmidt, F. L., Berner, J. G., & Hunter, J. E. (1973). Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 58, 5-9.

Smith, M. L. (1980). Publication bias and meta-analysis. Evaluation in Education, 4, 22-24.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752-760.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). The benefits of psychotherapy. Baltimore: Johns Hopkins University Press.

Strube, M. J., & Hartmann, D. P. (1983). Meta-analysis: Techniques, applications, and functions. Journal of Consulting and Clinical Psychology, 51, 14-27.

Sudman, S., & Bradburn, N. M. (1974). Response effects in surveys: A review and synthesis. Chicago: Aldine.

Tippett, L. H. C. (1931). The methods of statistics. London: Williams & Norgate.

Underwood, B. J. (1957). Interference and forgetting. Psychological Review, 64, 49-60.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. Psychological Bulletin, 91, 461-481.

Wilkinson, B. (1951). A statistical consideration in psychological research. Psychological Bulletin, 48, 156-158.

Wilson, G. T., & Rachman, S. J. (1983). Meta-analysis and the evaluation of psychotherapy outcome: Limitations and liabilities. Journal of Consulting and Clinical Psychology, 51, 54-64.

Wolf, F. M., Savickas, M. L., Saltzman, G. A., & Walker, M. L. (1983). A meta-analytic evaluation of an interpersonal skills curriculum for medical students: Synthesizing evidence over successive occasions. Unpublished manuscript. (Preliminary version presented at the meeting of the American Educational Research Association, Montreal, April 1983.)

Wortman, P. M. (1983). Evaluation research: A methodological perspective. Annotated Review of Psychology, 34, 223-260.

Wortman, P. M., & Bryant, F. B. (in press). School desegregation and black achievement: An integrative review. Sociological Methods and Research.

Yates, F., & Cochran, W. G. (1938). The analysis of groups of experiments. Journal of Agricultural Science, 28, 556-580.